



Processamento de ficheiros em formato PDF

Clara Cunha

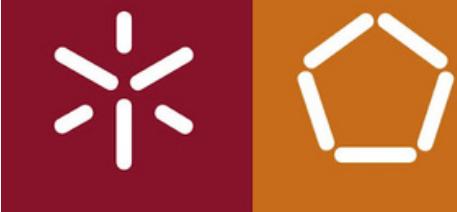
Manuel Carvalho

Nuno Matos



Documentos processados

- 01 glossario_neologismos_saude.pdf
- 02 diccionari-multilinguee-de-la-covid-19.pdf
- 03 glossario_de_termos_medicos_tecnicos_e_populares



Análise dos documentos

glossario_neologismos_saude.pdf

- **Termo**, seguido da sua **classe gramatical**

- **Equivalências linguísticas**

- **Definição**

- **Informações complementares** (opcionais):

- Siglas
- Notas enclopédicas
- Exemplos extraídos que ilustram o uso do termo.

degeneração macular relacionada à idade s.f.

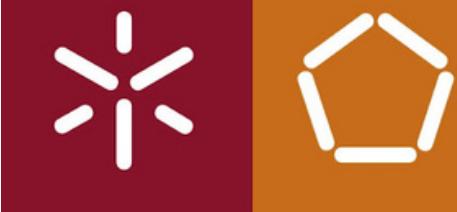
age-related macular degeneration [ing]; degeneración macular relacionada con la edad [esp]

Sigla: DMRI

Enfermidade que pode causar no ser humano a perda permanente da visão central.

Inf. encicl.: Apesar de não causar cegueira total, é uma das principais causas de perda visual em pessoas acima de 60 anos e por dificilmente afetar pessoas com menos de 60 anos é chamada de Relacionada à Idade.

“...A necessidade de tratar pacientes com Degeneração Macular Relacionada à Idade (DMRI), doença ocular que atinge, principalmente, pessoas acima dos 60 anos, motivou os pesquisadores do Instituto da Visão...” (162)



Tratamento dos dados

glossario_neologismos_saude.pdf

1. Conversão do formato PDF para TXT

```
pdftotext glossario_neologismos_saude.pdf neologismos.txt
```

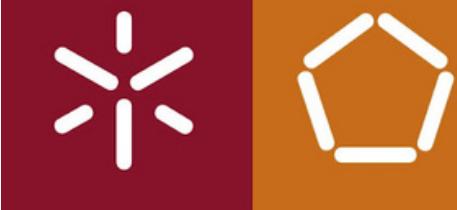
2. Leitura do ficheiro

```
with open('neologismos.txt', 'r', encoding='utf-8') as file:  
    content = file.read()
```

3. Extração da secção do glossário

```
pattern = r'FF3\.2(.*)?3\.3'  
match = re.search(pattern, content, re.DOTALL)
```

```
if match:  
    glossario_raw = match.group(1)[14:].strip()
```

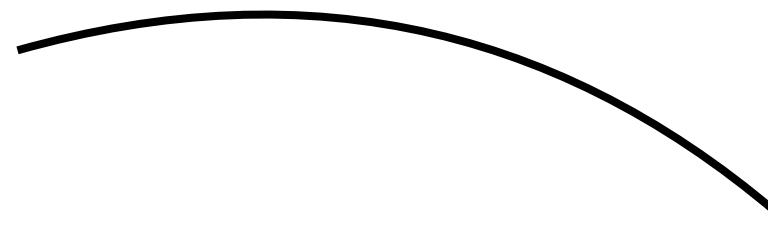


Tratamento dos dados

glossario_neologismos_saude.pdf

3. Extração da secção do glossário

```
1 AURI CLAUDIONEI MATOS FRÜBEL
2
3 GLOSSÁRIO DE NEOLOGISMOS TERMINOLÓGICOS
4 DA SAÚDE HUMANA: UMA CONTRIBUIÇÃO PARA A DESCRIÇÃO
5 DO LÉXICO CORRENTE DO PORTUGUÊS DO BRASIL
6
7 ARARAQUARA
8 2006
9
10 FF Frübel, Auri Claudionei Matos
11 Glossário de neologismos terminológicos da saúde humana:
12 uma contribuição para a descrição do léxico corrente do
13 português do Brasil / Auri Claudionei Matos Frübel - 2006
14 227 f.; 30 cm
15 Tese (Doutorado em Lingüística e Língua Portuguesa) -
16 Universidade Estadual Paulista, Faculdade de Ciências e Letras,
17 Campus de Araraquara
```



```
1 "abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode ser encontrada em todos os tipos de células do\organismo humano. Ao acumular-se excessivamente no córtex cerebral\ndo ser humano pode contribuir para o aceleramento do mal de alzheimer.\n"...
Pesquisadores alemães da Universidade de Bonn ajudaram a entender\ncomo a proteína abeta se acumula no córtex cerebral de portadores do\ncorpo de Alzheimer..." (159)\nação vasoconstritora s.f.\nvasoconstriction [ing]; acción vasoconstritora [esp]\nRedução do diâmetro das veias artérias do organismo humano, o que\nimplica na elevação da pressão sangüínea.\n"...\ndescobriram como atuam diferentes versões dos genes que controlam\nna produção de duas enzimas essenciais para a sobrevivência por fazer a\npressão arterial subir ou cair: a enzima conversora da angiotensina\n(ECA), que reduz o diâmetro das artérias (ação vasoconstritora) e eleva\nna pressão..." (180)\nacidente vascular cerebral isquêmico s.m.\nischemic cerebrovascular accident [ing];\nacidente vascular cerebral\nnisquêmico [esp]\nSigla: AVCI\nLesão que causa a morte de parte do cérebro humano, em virtude da falta\nde oxigênio e nutrientes que são inseridos no cérebro por meio da\ncirculação sangüínea.\nInf. encicl.: Uma vez privados do sangue, os neurônios morrem e liberam\nglutamato, uma substância química que realiza a comunicação entre as\ncélulas nervosas. Em concentrações elevadas, porém, o glutamato tornase tóxico e mata as células vizinhas, ampliando o estrago. O problema\npode levar à imobilidade de braços e pernas bem como causar a perda da\nfala. "...Conhecido como acidente vascular cerebral isquêmico (AVCI) ou\nnisquemia cerebral, esse problema pode levar à
```



Tratamento dos dados

glossario_neologismos_saude.pdf

4. Limpeza de dados

- Quebra de página

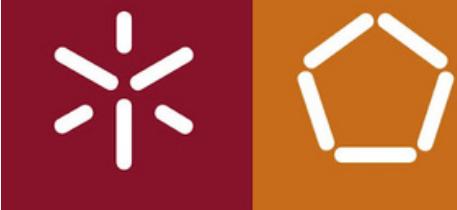
```
glossario_raw = re.sub(r'\nFF', ' ', glossario_raw)
```

- Nome de publicações

```
glossario_raw = re.sub(r'" [^"]+', " ", glossario_raw)
```

- Divisão de entradas

```
glossario_raw = re.sub(r'" [^"]+', " ", glossario_raw)
```

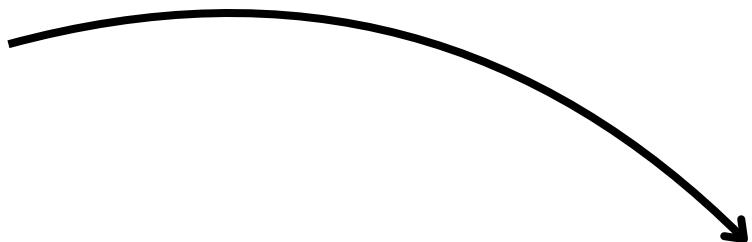


Tratamento dos dados

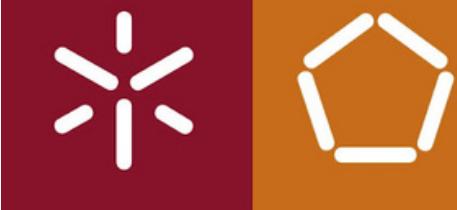
glossario_neologismos_saude.pdf

4. Limpeza de dados

```
1 "abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode ser encontrada em todos os tipos de células do\organismo humano. Ao acumular-se excessivamente no córtex cerebral\ndo ser humano pode contribuir para o aceleramento do mal de alzheimer.\n\"... Pesquisadores alemães da Universidade de Bonn ajudaram a entender\ncomo a proteína abeta se acumula no córtex cerebral de portadores do\mal de Alzheimer...\" (159)\nação vasoconstritora s.f.\nvasoconstriction [ing]; acción vasoconstritora [esp]\nRedução do diâmetro das veias artérias do organismo humano, o que\nimplica na elevação da pressão sangüínea.\n\"...descobriram como atuam diferentes versões dos genes que controlam\nna produção de duas enzimas essenciais para a sobrevivência por fazer a\npressão arterial subir ou cair: a enzima conversora da angiotensina\n(ECA), que reduz o diâmetro das artérias (ação vasoconstritora) e eleva\nna pressão...\" (180)\nacidente vascular cerebral isquêmico s.m.\nischemic cerebrovascular accident [ing]; acidente vascular cerebral\nisquêmico [esp]\nSigla: AVCI\nLesão que causa a morte de parte do cérebro humano, em virtude da falta\nde oxigênio e nutrientes que são inseridos no cérebro por meio da\ncirculação sangüínea.\nInf. encicl.: Uma vez privados do sangue, os neurônios morrem e liberam\nglutamato, uma substância química que realiza a comunicação entre as\ncélulas nervosas. Em concentrações elevadas, porém, o glutamato tornase tóxico e mata as células vizinhas, ampliando o estrago. O problema\npode levar à imobilidade de braços e pernas bem como causar a perda da\nfala. \"...Conhecido como acidente vascular cerebral isquêmico (AVCI) ou\nisquemia cerebral, esse problema pode levar à
```



```
"abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode ser encontrada em todos os\n159)\nação vasoconstritora s.f.\nvasoconstriction [ing]; acción vasoconstritora [esp]\n180)\nacidente vascular cerebral isquêmico s.m.\nischemic cerebrovascular accident\n37)\nnácido acetilsalicílico s.m.\nacetilsalicilic acid [ing]; ácido acetilsalicílico\n66)\nácido caínico s.m.\nkainic acid [ing]; ácido caínico [esp]\nSubstância utilizada\n09)\nnácido gama-aminobutírico s.m.\ngamma-aminobutyric acid [ing]; ácido gama-amino\n39, 136)\nnacidose metilmalônica s.f.\nmetilmalonic acidose [ing]; acidemia metilmalônica\n206)\nadjuvante genético s.m.\ngenetic adjuvant [ing]; adjuvante genético [esp]\nCo\n213)\nalergia asmática s.f.\nasthmatic allergy [ing]; alergia asmática [esp]\nHiper\n245)\nalfa-tocoferol s.m.\nalpha-tocopherol [ing]; alfa-tocoferol [esp]\nSubstância utilizada\n208)\nálumā s.m.\nvernonia condensata [ing]; alumá [esp]\nPlanta medicinal classificada\n21)\nnamiloidose sistêmica senil s.f.\nsenile systemic amyloidosis [ing]; amiloidose sistêmica\n117)\n análise biomecânica s.f.\nbiomechanic analysis [ing]; análise biomecânica [esp]\n80)\nnangina instável s.f.\nunstable angina [ing]; angina inestable [esp]\nAnomalia\n124)\nnangiografia de retina s.f.\nretina angiography [ing]; angiografia de la retina\n162)\nnantimonal pentavalente s.m.\npentavalent antimonial [ing]; antimonial pentavalente\n78)\napicoplasto s.f.\napicoplast [ing]; apicoplasto [esp]\nEstrutura responsável\n35)\napoptose s.f.\napoptose [ing]; apoptosis [esp]\nProcesso que denomina a morte programada
```



Tratamento dos dados

glossario_neologismos_saude.pdf

5. Extração da informação relevante

- Título

```
title_pattern = r'(.+) \w\.'
```

```
title = re.search(title_pattern, entry).group(1).strip()
```



```
[  
    "abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode  
    "159)\nAÇÃO vasoconstritora s.f.\nvasoconstriction [ing]  
    "180)\nacidente vascular cerebral isquêmico s.m.\nischemic  
    "37)\nácido acetilsalicílico s.m.\nacetilsalicylic acid
```

- Classe gramatical

```
gender_pattern = r'.+\w\.(\\w)\.'
```

```
gender = re.search(gender_pattern, entry).group(1)
```



```
[  
    "abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode  
    "159)\nAÇÃO vasoconstritora s.f.\nvasoconstriction [ing]  
    "180)\nacidente vascular cerebral isquêmico s.m.\nischemic  
    "37)\nácido acetilsalicílico s.m.\nacetilsalicylic acid
```

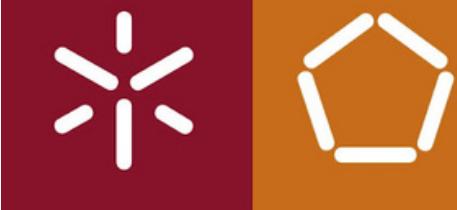
- Traduções

```
eng_trans = re.search(r'(.*)\[ing\]', entry)
```

```
esp_trans = re.search(r';(.*)\[esp\]', entry, flags=re.DOTALL)
```



```
[  
    "abeta s.f.\nabeta [ing]; abeta [esp]\nProteína que pode ser e  
    "159)\nAÇÃO vasoconstritora s.f.\nvasoconstriction [ing]; acci  
    "180)\nacidente vascular cerebral isquêmico s.m.\nischemic cer  
    "37)\nácido acetilsalicílico s.m.\nacetilsalicylic acid [ing];
```



Tratamento dos dados

glossario_neologismos_saude.pdf

5. Extração da informação relevante

- Descrição

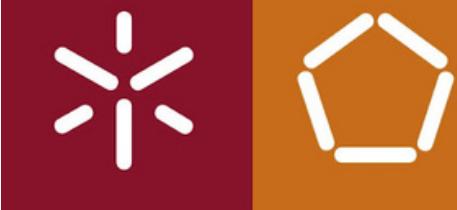
```
definition_pattern = r'\]\n([\s\S]*?\.)'
definition = re.search(definition_pattern, entry)
definition = definition.group(1) if definition else ""
```

- Sigla

```
sigla = re.search(r"Sigla: (.*)\n", definition)
if sigla:
    sigla = sigla.group(1).strip()
    definition = re.sub(r"Sigla: (.*)\n", "", definition)
```



```
\nSigla: AVCI\nLesão que causa a morte de parte do cérebro humano,
r humano, como prevenção ao\nenfarte, desde que o usuário não possu
```



Tratamento dos dados

glossario_neologismos_saude.pdf

5. Extração da informação relevante

- Referência encyclopédica

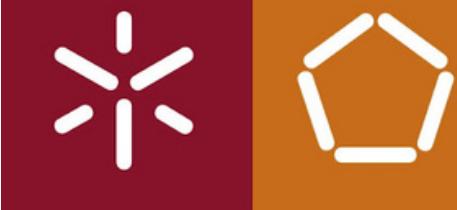
```
encicl = re.search(r"Inf. encicl.: (.*\n)", entry, flags=re.DOTALL)
if encicl:
    encicl = encicl.group(1).replace("\n", " ").strip()
```

```
In [1]: Inf. encicl.: Uma vez privados do sangue, os neurônios
      s de 30%, mas isso só vale para 75% da população...,
```

- Exemplo de uso

```
usage_pattern = r"^(?![\n])*(?=[\n])"
usage = re.search(usage_pattern, entry, flags=re.DOTALL)
```

```
In [1]: ibuir para o aceleramento do mal de alzheimer.
      ...Pesquisadores alemães da Universida
      ...descobriram como atuam diferentes versões dos genes que controlam na produção de d
      anó, em virtude da falta de oxigênio e nutrientes que são inseridos no cérebro por mei
      possua altos níveis de colesterol.
      ...A aspirina (ácido acetilsalicílico) pode reduzi
      o experimento, eles aplicaram a mesma quantidade de três compostos químicos capazes de
      humano, controlando, na célula, a entrada de partículas de carga elétrica negativa,
      o retardamento do desenvolvimento infantil.
      ...a acidose metilmalônica, doença que
```



Tratamento dos dados

glossario_neologismos_saude.pdf

5. Extração da informação relevante

- Organização das informações num dicionário

```
content = {
    'traducoes': {'en': eng_trans.group(1).replace("\n", " ").strip() if eng_trans else "",
                  'es': esp_trans.group(1).replace("\n", " ").strip() if esp_trans else ""},
    'categoria': "n " + gender,
    'descricao': definition.replace("\n", " ").strip(),
    'exemplo': usage.group(1).replace("\n", " ").strip() if usage else ""
}
```

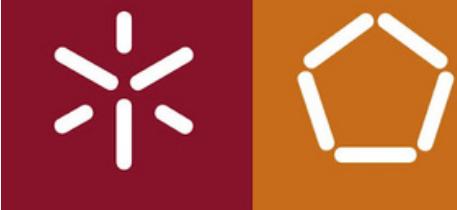
```
if sigla:
```

```
    content['sigla'] = sigla
```

```
if encicl:
```

```
    content['enciclopedia'] = encicl
```

```
for conceito in glossario[:-1]:
    title, content = parse_entry(conceito)
    dict[title] = content
```



Tratamento dos dados

glossario_neologismos_saude.pdf

6. Intervenções manuais

- Linha 3672

“es” → “[esp]”

gastric cancer [ing]; câncer gástrico [es]

- Linhas 6109, 4678 e 3420

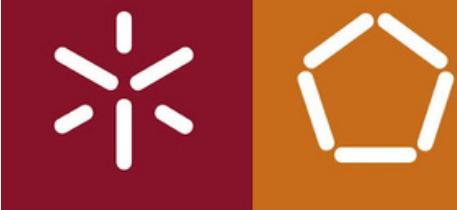
Inserção do caractere de aspas
no exemplo

“...A autora reporta-se também à crescente complexidade da biologia sistêmica, que mapea redes de sinais de informações que se entrelaçam por meio de um número nunca antes imaginado de ligações específicas de proteínas, resultando em respostas finais temporalmente e mecanisticamente precisas... (17)

- Linha 3337

‘ → “

“...Certeza mesmo, só na autópsia, quando o cérebro revela seu terrível estado de degradação: tecido cerebral atrofiado, fibras nervosas emaranhadas, neurônios invadidos por placas da proteína betaamilóide...’ (97, 136, 209)

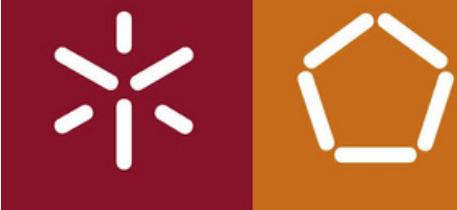


Tratamento dos dados

glossario_neologismos_saude.pdf

RESULTADO

```
"ação vasoconstritora": {
    "traducoes": {
        "ingles": "vasoconstriction",
        "espanhol": "acción vasoconstritora"
    },
    "genero": "f",
    "descricao": "Redução do diâmetro das veias artérias do organismo humano, o que implica na elevação da pressão sanguínea",
    "exemplo": "...descobriram como atuam diferentes versões dos genes que controlam a produção de duas enzimas essenciais
},
"acidente vascular cerebral isquêmico": {
    "traducoes": {
        "ingles": "ischemic cerebrovascular accident",
        "espanhol": "accidente vascular cerebral isquêmico"
    },
    "genero": "m",
    "descricao": "Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigênio e nutrientes que são
    "exemplo": "...Conhecido como acidente vascular cerebral isquêmico (AVCI) ou isquemia cerebral, esse problema pode lev
    "sigla": "AVCI",
    "enciclopedia": "Uma vez privados do sangue, os neurônios morrem e liberam glutamato, uma substância química que reali
```



Análise dos documentos

diccionari-multilinguee-de-la-covid-19.pdf

- Número de ordem
- Termo em catalão
- Variantes
- Sinónimos complementares
- Equivalências linguísticas
- Área temática
- Definição
- Notas adicionais

2 acalabrutinib n m
oc acalabrutinib n m
eu akalabrutinib n
gl acalabrutinib n m
es acalabrutinib n m
en acalabrutinib n m
fr acalabrutinib n m
pt [PT] acalabrutinib n m
pt [BR] acalabrutinibe n m
nl acalabrutinib n
ar اکالابرتوینیب
CAS 1420477-60-6
PRINCIPI ACTIUS. Fàrmac antineoplàstic que bloca la tirosina-cinasa de Bruton i inhibeix la replicació dels limfòcits T cancerosos.
Nota: 1. L'acalabrutinib s'empra en el tractament de la leucèmia limfocítica crònica i de diversos tipus de limfomes. També s'investiga per a tractar altres tipus de càncer. Se n'ha suggerit l'ús per al tractament de la COVID-19. És d'origen sintètic.
2. La denominació acalabrutinib és la forma catalana corresponent a la DCI.



Tratamento dos dados

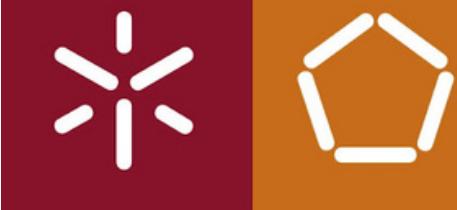
diccionari-multilinguee-de-la-covid-19.pdf

1. Conversão do formato PDF para xml

```
pdftohtml -xml -i .\diccionari-multilinguee-de-la-covid-19.pdf multilingue.xml
```

2. Leitura do ficheiro

```
with open('multilingue.xml', 'r', encoding='utf-8') as file:  
    content = file.read()
```



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

3. Extração das páginas

```
for page in soup.find_all('page'):
    page_number = int(page['number'])

    if page_number % 2 == 0:
        left = 500
    else:
        left = 440

    if 30 <= page_number <= 182:
        left_column = ""
        right_column = ""

        for text in page.find_all('text'):
            if int(text['left']) < left:
                left_column += str(text) + '\n'
            else:
                right_column += str(text) + '\n'

        page = left_column + right_column
        pages += (page)
```

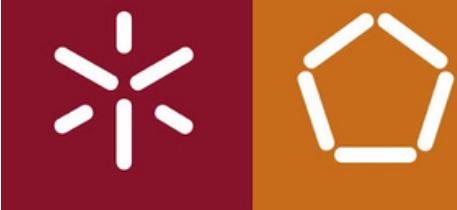


Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

4. Limpeza de dados

```
pages = re.sub(r'<text  
font="6".*/text>\n', r'', pages)  
pages = re.sub(r'<.*font="14".*/text>\n',  
r'', pages)  
pages = re.sub(r'<.*font="15".*/text>\n',  
r'', pages)  
pages = re.sub(r'<.*font="36".*/text>\n',  
r'', pages)  
pages = re.sub(r'<.*font="38".*/text>\n',  
r'', pages)  
pages = re.sub(r'<.*?>', r'', pages)
```

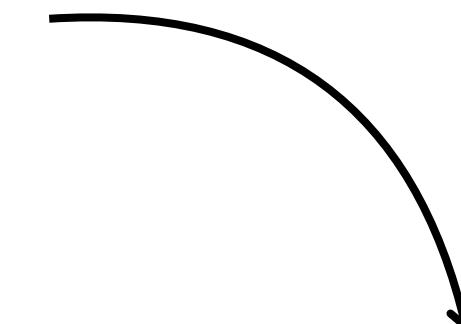


Tratamiento dos dados

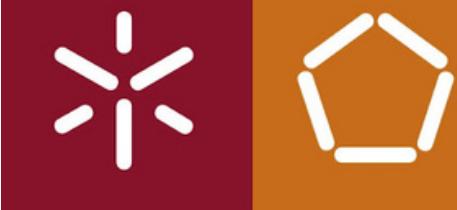
diccionari-multilinguee-de-la-covid-19.pdf

4. Limpeza de dados

```
"<text font=\\"6\\" height=\\"25\\" left=\\"85\\" top=\\"1149\\" width=\\"24\\">28</text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"355\\" width=\\"11\\">1 </text>\n<text font=\\"25\\" height=\\"16\\" left=\\"157\\" top=\\"354\\" width=\\"28\\"><b>ACA</b></text>\n<text font=\\"12\\" height=\\"16\\" left=\\"185\\" top=\\"355\\" width=\\"26\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"376\\" width=\\"94\\"> veg. </text>\n<text font=\\"25\\" height=\\"16\\" left=\\"187\\" top=\\"375\\" width=\\"113\\"><b>assaig aleatoritzat</b></text>\n<text font=\\"12\\" height=\\"16\\" left=\\"299\\" top=\\"376\\" width=\\"26\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"412\\" width=\\"11\\">2 </text>\n<text font=\\"25\\" height=\\"16\\" left=\\"157\\" top=\\"411\\" width=\\"82\\"><b>acalabrutinib</b></text>\n<text font=\\"12\\" height=\\"16\\" left=\\"239\\" top=\\"412\\" width=\\"26\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"437\\" width=\\"3\\"> </text>\n<text font=\\"12\\" height=\\"16\\" left=\\"157\\" top=\\"437\\" width=\\"18\\"><i>oc </i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"187\\" top=\\"437\\" width=\\"81\\">acalabrutinib</text>\n<text font=\\"12\\" height=\\"16\\" left=\\"268\\" top=\\"437\\" width=\\"26\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"458\\" width=\\"3\\"> </text>\n<text font=\\"12\\" height=\\"16\\" left=\\"157\\" top=\\"458\\" width=\\"19\\"><i>eu </i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"187\\" top=\\"458\\" width=\\"81\\">akalabrutinib</text>\n<text font=\\"12\\" height=\\"16\\" left=\\"268\\" top=\\"458\\" width=\\"11\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"479\\" width=\\"3\\"> </text>\n<text font=\\"12\\" height=\\"16\\" left=\\"157\\" top=\\"479\\" width=\\"15\\"><i>gl </i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"187\\" top=\\"479\\" width=\\"81\\">acalabrutinib</text>\n<text font=\\"12\\" height=\\"16\\" left=\\"268\\" top=\\"479\\" width=\\"26\\"><i> n m</i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"123\\" top=\\"500\\" width=\\"3\\"> </text>\n<text font=\\"12\\" height=\\"16\\" left=\\"157\\" top=\\"500\\" width=\\"17\\"><i>es </i></text>\n<text font=\\"11\\" height=\\"16\\" left=\\"187\\" top=\\"500\\" width=\\"81\\">acalabrutinib</text>\n<text font=\\"12\\" height=\\"16\\" left=\\"268\\" top=\\"500\\" width=\\"268\\">
```



"1 \nACA\n n m\n veg. \nassaig aleatoritzat\n n m\n2 \nacalabrutinib\n n m\n \noc \nacalabrutinib\n n m\n \neu \nakalabrutinib\n n n \ngl \nacalabrutinib\n n m\n \nes \nacalabrutinib\n n m\n \nen \nacalabrutinib\n n m\n \nfr \nacalabrutinib\n n m\n \npt [PT] \nacalabrutinib\n n m\n \npt [BR] \nacalabrutinibe\n n m\n \nnl \nacalabrutinib\n n n \nar \nبروتوكول انتيبيوتيرال\nnCAS \n1420477-60-6\nPRINCIPIS ACTIUS. Fàrmac antineoplàstic que bloca la tirosina-cinasa de Bruton i inhibeix la replicació dels limfòcits T cancerosos.\nNota: 1. L'acalabrutinib s'empra en el tractament de la leucèmia limfocítica crònica i de diversos tipus de limfomes. També s'investiga per a tractar altres tipus de càncer. Se n'ha suggerit l'ús per al tractament de la COVID-19. És d'origen sintètic.\n2. La denominació \nacalabrutinib\n és la forma catalana \ncorrespondent a la DCI.\n3 \nàcid desoxiribonucleic\n n m\nnsigla \nADN\n n m\n \noc \nacid desoxiribonucleic\n n m\n \neu \nazido desoxirribonukleiko\n n n; DNA\n n n \ngl \nácido desoxirribonucleico\n n m\n; ADN\n n m\n; \nDNA\n n m\n \nes \nácido desoxirribonucleico\n n m\n; ADN\n n m\n; \nDNA\n n m\n \nen \ndeoxyribonucleic acid\n n n; DNA\n n n \nfr \nacide désoxyribonucléique\n n m\n; DNA\n n m\n; \nADN\n n m\n \npt \nácido desoxirribonucleico\n n m\n; ADN\n n m\n; \nDNA\n n m\n \nnl \ndesoxyribonucleïneuur\n n n; DNA\n n n \nar \nبروتوكول انتيبيوتيرال صوقنم يبير يوون ضمح \nETIOPATOGÈNIA. Àcid nucleic constituit per \nnucleòtids de desoxiribosa, àcid fosfòric i les bases \nnitrogenades adenina, citosina, guanina i timina, que es troba fonamentalment en el nucli, en els mitocondris i en els cloroplasts, i que constitueix la base molecular de l'erència biològica.\nNota: 1. La sigla \nADN\n nté un ús divulgatiu, mentre que en \nàmbits especialitzats se sol utilitzar la sigla anglesa \nDNA\n. \nAquesta recomanació és aplicable també a la parella de sigles \nARN-RNA\n, atès que són les formes clarament identificables dins la comunitat científica internacional. \nAnàlogament, les sigles creades a partir d'aquestes formes segueixen preferentment l'ordre internacional \n(\nmtDNA, rDNA, tRNA\n...), i només secundàriament i en un àmbit divulgatiu es formen a partir de l'ordre romànic, \namb \nADN\n i \nARN \n(\nADNm, ADNr, ARNt\n...). \n2. La sigla \nDNA\n correspon a l'anglès \ndeoxyribonucleic \nacid \n



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

5. Marcações

markers = "@#§★∞⌚฿\$"

```
categorias = {  
    r"n",  
    r"n pl",  
    r"n m",  
    r"n m pl",  
    r"n f",  
    r"n f pl",  
    r"n m, f",  
    r"n m/f",  
    r"adj",  
    r"v tr",  
    r"v tr/intr",  
    r"v intr"  
}
```

```
remis = {  
    r'sin. compl.',  
    r'sin.',  
    r'den. com.',  
    r"sigla",  
    r"veg."  
}
```

```
langs = {  
    r"oc",  
    r"eu",  
    r"gl",  
    r"es",  
    r"en",  
    r"fr",  
    r"pt",  
    r"pt \[PT\]",  
    r"pt \[BR\]",  
    r"nl",  
    r"ar"  
}
```

```
codigos = {  
    r"sbl",  
    r"nc",  
    r"CAS"  
}
```

```
campos = {  
    r"CONCEPTES  
GENERALS",  
    r"EPIDEMIOLOGIA",  
    r"ETIOPATOLOGÈNIA",  
    r"DIAGNÒSTIC",  
    r"CLÍNICA",  
    r"PREVENCIÓ",  
    r"TRACTAMENT",  
    r"PRINCIPIIS ACTIUS",  
    r"ENTORN SOCIAL"  
}
```



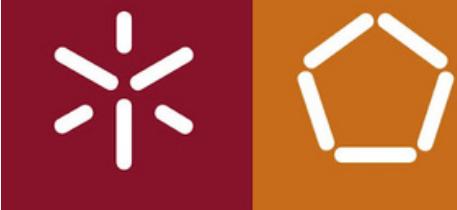
Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

5. Marcações

- Categorias

```
new_pages = "
for line in pages.split('\n'):
    line = line.strip()
    if line in categorias:
        new_pages += '##' + line + '\n' # marcar as categorias com ##
    else:
        new_pages += line + '\n'
```



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

5. Marcações

- "CAS", "sbl" e "nc"

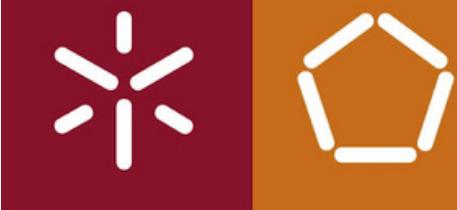
```
codigos_pattern = rf"^(?=\^{markers})]"  
pages = re.sub(codigos_pattern, r'(marcador referido acima)\l\n', pages, flags=re.MULTILINE)
```

- "en", "pt" e "es"

```
traducao_pattern = rf"^(?=\^{markers})]"  
pages = re.sub(traducao_pattern, r'★\l\n', pages, flags=re.MULTILINE)
```

- "sigla", "sin" (sinónimo) e "veg" (ver)

```
remis_pattern = rf"^(?=\^{markers})]"  
pages = re.sub(remis_pattern, r'∞\l\n', pages, flags=re.MULTILINE)
```



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

5. Marcações

- Diferentes tipos de descrição

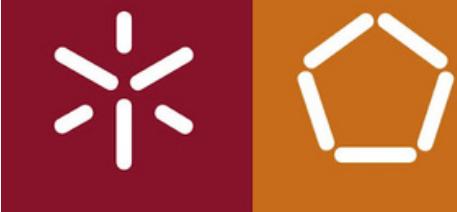
```
campos_pattern = rf"^(?={'.join(campos)})(?=[^{markers}])"
pages = re.sub(campos_pattern, r'@\\1.', pages, flags=re.MULTILINE)
```

- Notas

```
nota_pattern = r"^(Nota:.*)"
pages = re.sub(nota_pattern, r'@\\1', pages, flags=re.MULTILINE)
```

- Identificador único

```
concept_pattern = r"([1-9]\\d*(\\n+.*){1,2}\\n##.+?)"
pages = re.sub(rf"{concept_pattern}", r"@\\1", pages, flags= re.MULTILINE)
```

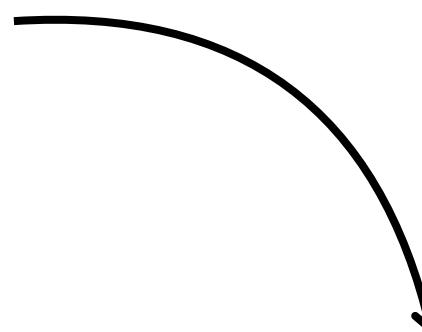


Tratamiento dos dados

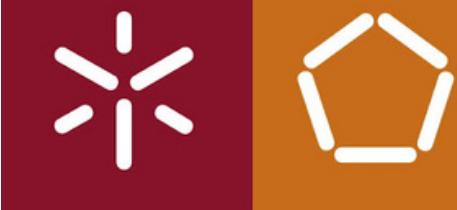
diccionari-multilinguee-de-la-covid-19.pdf

5. Marcações

"1 \nACA\n n m\n veg. \nassaig aleatoritzat\n n m\n2 \nacalabrutinib\n n m\n \noc \nacalabrutinib\n n m\n \neu \nakalabrutinib\n n\n \ngl \nacalabrutinib\n n m\n \nes \nacalabrutinib\n n m\n \nen \nacalabrutinib\n n m\n \nfr \nacalabrutinib\n n m\n \npt [PT] \nacalabrutinib\n n m\n \npt [BR] \nacalabrutinibe\n n m\n \nnl \nacalabrutinib\n n\n \nar \nبنیتوبیدا\ncas \n1420477-60-6\nPRINCIPIS ACTIUS. Fàrmac antineoplàstic que bloca \nla tirosina-cinasa de Bruton i inhibeix la replicació \nndels limfòcits T cancerosos.\nNota: 1. L'acalabrutinib s'empra en el tractament \nde la leucèmia limfocítica crònica i de diversos tipus \nde limfomes. També s'investiga per a tractar altres \nntipus de càncer. Se n'ha suggerit l'ús per al tractament \nde la COVID-19. És d'origen sintètic.\n2. La denominació \nacalabrutinib\n és la forma catalana \nncorrespondent a la DCI.\n3 \nàcid desoxiribonucleic\n n m\nnsigla \nADN\n n m\n \noc \nàcid desoxiribonucleic\n n m\n \neu \nazido desoxirribonukleiko\n n; DNA\n n m\n \ngl \nàcido desoxirribonucleico\n n m\n; ADN\n n m\n \nnDNA\n n m\n \nes \nàcido desoxirribonucleico\n n m\n; ADN\n n m\n \nnDNA\n n m\n \nen \ndeoxyribonucleic acid\n n\n; DNA\n n m\n \nfr \nacide désoxyribonucléique\n n m\n; DNA\n n m\n; \nADN\n n m\n \npt \nàcido desoxirribonucleico\n n m\n; ADN\n n m\n \nnDNA\n n m\n \nnl \ndesoxyribonucleïneuur\n n\n; DNA\n n m\n \nar \nنحوه\نیجسکا\ صوقنم بیبر یوون ضمچا\ن ETIOPATOGÈNIA. Àcid nucleic constituït per \nnucleòtids de desoxiribosa, àcid fosfòric i les bases \nnitrogenades adenina, citosina, guanina i timina, \nque es troba fonamentalment en el nucli, en els \nmitocondris i en els cloroplasts, i que constitueix la \nbase molecular de l'herència biològica.\nNota: 1. La sigla \nADN\n nté un ús divulgatiu, mentre que en \nàmbits especialitzats se sol utilitzar la sigla anglesa\n nDNA\n. \nAquesta recomanació és aplicable també a la parella \nde sigles \nARN-RNA\n, atès que són les formes clarament \nidentificables dins la comunitat científica internacional. \nAnàlogament, les sigles creades a partir d'aquestes \nformes segueixen preferentment l'ordre internacional \n(\nmtDNA, rDNA, tRNA\n...), i només secundàriament i en un \nàmbit divulgatiu es formen a partir de l'ordre romànic, \namb \nADN\n i \nARN \n(\nADNm, ADNr, ARNt\n...). \n2. La sigla \nDNA\n correspon a l'anglès \ndeoxyribonucleic \nacid \n



"@1\nACA\n## m\nveg. \nassaig aleatoritzat\n## m\n@2\nacalabrutinib\n## n\n\n★oc\nacalabrutinib\n## n\n\n★eu\nakalabrutinib\n## n\n\n★gl\nacalabrutinib\n## m\n\n★es\nacalabrutinib\n## m\n\n★en\nacalabrutinib\n## m\n\n★fr\nacalabrutinib\n## m\n\n★pt [PT]\nacalabrutinib\n## m\n\n★pt [BR]\nacalabrutinibe\n## m\n\n★nl\nacalabrutinib\n## n\n\n★ar \nبنیتوبیدا\ncas \n1420477-60-6\nPRINCIPIS ACTIUS\\. Fàrmac antineoplàstic que bloca \nla tirosina-cinasa de Bruton i inhibeix la replicació \nndels limfòcits T cancerosos.\nNota: 1. L'acalabrutinib s'empra en el tractament \nde la leucèmia limfocítica crònica i de diversos tipus \nde limfomes. També s'investiga per a tractar altres \nntipus de càncer. Se n'ha suggerit l'ús per al tractament \nde la COVID-19. És d'origen sintètic.\n2. La denominació \nacalabrutinib\n és la forma catalana \nncorrespondent a la DCI.\n3 \nàcid desoxiribonucleic\n## m\nnsigla \nADN\n## m\n\n★oc\nàcid desoxiribonucleic\n## m\n\n★eu\nazido desoxirribonukleiko\n n; DNA\n## n\n\n★gl \nàcido desoxirribonucleico\n## m\n; ADN\n## m\n\nnDNA\n## m\n\n★es \nàcido desoxirribonucleico\n## m\n; ADN\n## m\n\nnpt \nàcido desoxirribonucleico\n## m\n; ADN\n## m\n\nnDNA\n## m\n\nn★en \ndeoxyribonucleic acid\n## n\n; DNA\n## n\n\n★fr \nacide désoxyribonucléique\n## m\n; DNA\n## m\n; \nADN\n## m\n\nnpt \nàcido desoxirribonucleico\n## m\n; ADN\n## m\n\nnDNA\n## m\n\nn★nl \ndesoxyribonucleïneuur\n## n\n; DNA\n## m\n\nn★ar \nنحوه\نیجسکا\ صوقنم بیبر یوون ضمچا\n ETIOPATOGÈNIA\\. Àcid nucleic constituït per \nnucleòtids de desoxiribosa, àcid fosfòric i les bases \nnitrogenades adenina, citosina, guanina i timina, \nque es troba fonamentalment en el nucli, en els \nmitocondris i en els cloroplasts, i que constitueix la \nbase molecular de l'herència biològica.\nNota: 1. La sigla \nADN\n nté un ús divulgatiu, mentre que en \nàmbits especialitzats se sol utilitzar la sigla anglesa\n nDNA\n. \nAquesta recomanació és aplicable també a la parella \nde sigles \nARN-RNA\n, atès que són les formes clarament \nidentificables dins la comunitat científica internacional.\nAnàlogament, les sigles creades a partir d'aquestes \nformes segueixen preferentment l'ordre internacional \n(\nmtDNA, rDNA, tRNA\n...), i només secundàriament i en un \nàmbit divulgatiu es formen a partir de l'ordre romànic, \namb \nADN\n i \nARN \n(\nADNm, ADNr, ARNt\n...). \n2. La sigla \nDNA\n correspon a l'anglès \ndeoxyribonucleic \nacid \n(\n'n\nàcid desoxiribonucleic\n'). \n@4\nn\nàcid ribonucleic\n## m\n\n★sigla\nARN\n## m\n\n★en\nRNA\n## m\n\n★eu\nazido erribonukleiko\n## m\n; RNA\n## n\n\n★gl \nàcido ribonucleico\n## m\n; ARN\n## m\n\nn\n★es \nàcido ribonucleico\n## m\n; ARN\n## m\n\nn\n★fr \nacide ribonucléique\n## m\n; ARN\n## m\n\nn\n★pt \nàcido ribonucleico\n## m\n; ARN\n## m\n\nn\n★ar \nنحوه\نیجسکا\ صوقنم بیبر یوون ضمچا\n ETIOPATOGÈNIA\\. Àcid nucleic constituït per \nnucleòtids de ribosa, àcid fosfòric i les bases \nnitrogenades adenina, citosina, guanina i uracil, \nque es troba fonamentalment en el nucli,



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

6. Extração da Informação Relevante

- Título

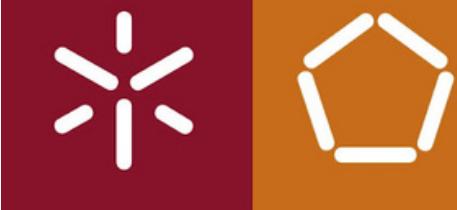
```
title_pattern = r"^\d+((\n.*){1,2})\n##(.*)"  
title = re.search(title_pattern, concept, flags=re.MULTILINE)
```

- Categoria

```
categoria_pattern = r"^\d+((\n.*){1,2})\n##(.*)"  
categoria = re.search(categoria_pattern, concept, flags=re.MULTILINE)
```

- Traduções

```
traducoes_pattern = rf"★((.*\n)*?)(@|§|∞|⌚|$)"  
traduc = re.findall(traducoes_pattern, concept, flags=re.MULTILINE)
```



Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

6. Extração da Informação Relevante

- Remissões

```
remi_pattern = rf"∞((.*\n)*?)(@|§|★|⌚|฿|$)"  
remiss = re.findall(remi_pattern, concept, flags=re.MULTILINE)
```

- Códigos

```
codigos_pattern = rf"§((.*\n)*?)(?= {'|'.join(markers)})"  
codigos = re.findall(codigos_pattern, concept, flags=re.MULTILINE)
```

- Descrição

```
campos_pattern = rf"⌚.*\.(.*\n)*?)(?= {'|'.join(markers)})"  
campos = re.search(campos_pattern, concept, flags=re.MULTILINE)
```



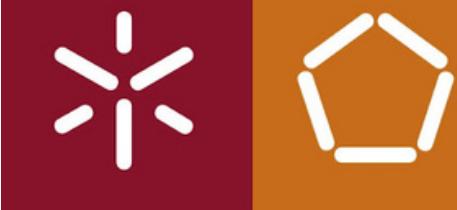
Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

6. Extração da Informação Relevante

- Notas

```
nota_pattern = rf"\\Nota:(.*\\n)*?)(?=\\{'.join(markers)\\})"  
notas = re.search(nota_pattern, concept, flags=re.MULTILINE)
```

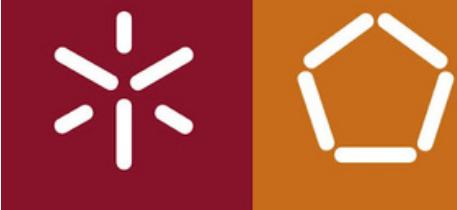


Tratamento dos dados

diccionari-multilinguee-de-la-covid-19.pdf

RESULTADO

```
"adequacio de l'estorç terapeutic": {
    "categoria": "n f",
    "traducoes": {},
    "remis": {
        "veg.": {
            "texto": "adequació de les actuacions sanitàries",
            "categoria": "n f"
        }
    },
    "codigos": {}
},
"adequació de les actuacions sanitàries": {
    "categoria": "n f",
    "traducoes": {
        "oc": {
            "1": {
                "tradução": "adequacion des actuacions sanitàries",
                "categoria": "n f"
            },
            "2": {
                "tradução": "limitacion der esfòrç terapeutic",
                "categoria": "n f"
            }
        },
        "eu": {
            "1": {
                "tradução": "ahalegin terapeutikoa egokitze",
                "categoria": "n"
            },
            "2": {
                "tradução": "ahaleginterapeutikoa mugatze",
                "categoria": "n"
            }
        }
    }
}
```



Análise dos documentos

Glossário de Termos Médicos Técnicos e Populares.pdf

- **Termo**

A

a milionésima parte de um grama (pop) , micrograma

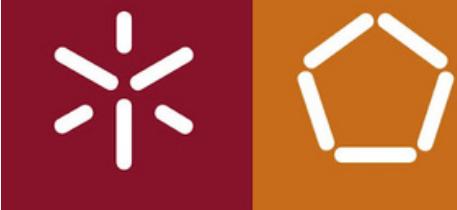
à volta da boca (pop) , perioral

à volta da órbita (pop) , periorbital

à volta dos vasos sanguíneos (pop) , perivascular

abaixamento, abatimento, prostração (pop) , depressão

- **Definição**



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

1. Conversão do formato PDF para XML

```
pdftohtml -xml Glossário de Termos Médicos Técnicos e Populares.pdf populares.xml
```

2. Leitura do ficheiro

```
with open('neologismos.txt', 'r', encoding='utf-8') as file:  
    content = file.read()
```

3. Extração da secção do glossário

```
start_pattern = r"<text[^>]*><b>A</b></text>"  
start_match = re.search(start_pattern, content)  
content = content[start_match.start():] if  
start_match else ""
```

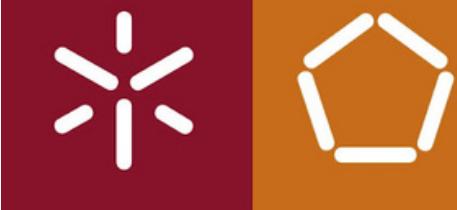


Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

3. Extração da secção do glossário

```
"<text top=\"405\" left=\"128\" width=\"30\" height=\"47\" font=\"4\"><b>A</b></text>\n<text top=\"474\" left=\"128\" width=\"246\" height=\"18\" font=\"5\"><i>a  
milionésima parte de um grama</i></text>\n<text top=\"474\" left=\"374\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"474\" left=\"431\"  
width=\"95\" height=\"18\" font=\"1\"><b>micrograma</b></text>\n<text top=\"474\" left=\"526\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"515\"  
left=\"128\" width=\"112\" height=\"18\" font=\"5\"><i>à volta da boca</i></text>\n<text top=\"515\" left=\"240\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"515\" left=\"297\" width=\"61\" height=\"18\" font=\"1\"><b>perioral</b></text>\n<text top=\"515\" left=\"358\" width=\"5\" height=\"18\"  
font=\"2\"> </text>\n<text top=\"555\" left=\"128\" width=\"118\" height=\"18\" font=\"5\"><i>à volta da órbita</i></text>\n<text top=\"555\" left=\"246\"  
width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"555\" left=\"303\" width=\"81\" height=\"18\" font=\"1\"><b>periorbital</b></text>\n<text  
top=\"555\" left=\"384\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"596\" left=\"128\" width=\"218\" height=\"18\" font=\"5\"><i>à volta dos vasos  
sanguíneos</i></text>\n<text top=\"596\" left=\"346\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"596\" left=\"403\" width=\"97\"  
height=\"18\" font=\"1\"><b>perivasicular</b></text>\n<text top=\"596\" left=\"501\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"636\" left=\"128\"  
width=\"273\" height=\"18\" font=\"5\"><i>abaixamento, abatimento, prostração</i></text>\n<text top=\"636\" left=\"401\" width=\"57\" height=\"18\" font=\"2\">  
(pop) , </text>\n<text top=\"636\" left=\"459\" width=\"82\" height=\"18\" font=\"1\"><b>depressão</b></text>\n<text top=\"636\" left=\"541\" width=\"5\"  
height=\"18\" font=\"2\"> </text>\n<text top=\"677\" left=\"128\" width=\"66\" height=\"18\" font=\"1\"><b>abcesso</b></text>\n<text top=\"677\" left=\"194\"  
width=\"14\" height=\"18\" font=\"2\"> , </text>\n<text top=\"677\" left=\"208\" width=\"113\" height=\"18\" font=\"5\"><i>abcesso, tumor</i></text>\n<text  
top=\"677\" left=\"321\" width=\"48\" height=\"18\" font=\"2\"> (pop) </text>\n<text top=\"717\" left=\"128\" width=\"113\" height=\"18\" font=\"5\"><i>abcesso,  
tumor</i></text>\n<text top=\"717\" left=\"241\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"717\" left=\"299\" width=\"66\" height=\"18\"  
font=\"1\"><b>abcesso</b></text>\n<text top=\"717\" left=\"365\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"758\" left=\"128\" width=\"214\"  
height=\"18\" font=\"5\"><i>abcesso; acumulação de pus</i></text>\n<text top=\"758\" left=\"342\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text  
top=\"758\" left=\"399\" width=\"72\" height=\"18\" font=\"1\"><b>empíema</b></text>\n<text top=\"758\" left=\"471\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"798\" left=\"128\" width=\"73\" height=\"18\" font=\"1\"><b>abdômen</b></text>\n<text top=\"798\" left=\"201\" width=\"14\" height=\"18\"  
font=\"2\"> , </text>\n<text top=\"798\" left=\"215\" width=\"107\" height=\"18\" font=\"5\"><i>barriga, ventre</i></text>\n<text top=\"798\" left=\"322\"  
width=\"48\" height=\"18\" font=\"2\"> (pop) </text>\n<text top=\"839\" left=\"128\" width=\"82\" height=\"18\" font=\"1\"><b>abdominal</b></text>\n<text  
top=\"839\" left=\"210\" width=\"14\" height=\"18\" font=\"2\"> , </text>\n<text top=\"839\" left=\"225\" width=\"49\" height=\"18\" font=\"5\"><i>ventral</i></text>\n<text top=\"839\" left=\"274\" width=\"48\" height=\"18\" font=\"2\"> (pop) </text>\n<text top=\"879\" left=\"128\" width=\"75\" height=\"18\"  
font=\"1\"><b>aberrante</b></text>\n<text top=\"879\" left=\"203\" width=\"14\" height=\"18\" font=\"2\"> , </text>\n<text top=\"879\" left=\"217\" width=\"60\"  
height=\"18\" font=\"5\"><i>anormal</i></text>\n<text top=\"879\" left=\"277\" width=\"48\" height=\"18\" font=\"2\"> (pop) </text>\n<text top=\"920\" left=\"128\"  
width=\"120\" height=\"18\" font=\"5\"><i>abertura; orifício</i></text>\n<text top=\"920\" left=\"248\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>
```



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

4. Limpeza de dados

- Remoção de tags estruturais e de linhas com letras maiúsculas

```
tags = [
    r"</?fontspec.*?>",
    r"</?page.*?>",
    r"</?text.*?>",
    r"<i>",
    r"</i>"
]
content = re.sub(r"<.*>[A-Z]<.*>", "", content)
# letras
for tag in tags:
    content = re.sub(tag, "", content)
```

Tratamento dos dados



Glossário de Termos Médicos Técnicos e Populares.pdf

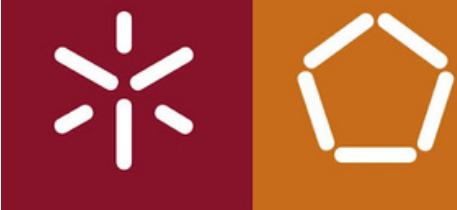
4. Limpeza de dados

- Remoção de tags estruturais e de linhas com letras maiúsculas

"<text top=\"405\" left=\"128\" width=\"30\" height=\"47\" font=\"4\">A</text>\n<text top=\"474\" left=\"128\" width=\"246\" height=\"18\" font=\"5\"><i>a milionésima parte de um grama</i></text>\n<text top=\"474\" left=\"374\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"474\" left=\"431\" width=\"95\" height=\"18\" font=\"1\">micrograma</text>\n<text top=\"474\" left=\"526\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"515\" left=\"128\" width=\"112\" height=\"18\" font=\"5\"><i>à volta da boca</i></text>\n<text top=\"515\" left=\"240\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"515\" left=\"297\" width=\"61\" height=\"18\" font=\"1\">perioral</text>\n<text top=\"515\" left=\"358\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"555\" left=\"128\" width=\"118\" height=\"18\" font=\"5\"><i>à volta da órbita</i></text>\n<text top=\"555\" left=\"246\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"555\" left=\"303\" width=\"81\" height=\"18\" font=\"1\">periorbital</text>\n<text top=\"555\" left=\"384\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"596\" left=\"128\" width=\"218\" height=\"18\" font=\"5\"><i>à volta dos vasos sanguíneos</i></text>\n<text top=\"596\" left=\"346\" width=\"57\" height=\"18\" font=\"2\"> (pop) , </text>\n<text top=\"596\" left=\"403\" width=\"97\" height=\"18\" font=\"1\">perivasicular</text>\n<text top=\"596\" left=\"501\" width=\"5\" height=\"18\" font=\"2\"> </text>\n<text top=\"636\" left=\"128\" width=\"273\" height=\"18\" font=\"5\"><i>abaixamento, abatimento, prostração</i></text>\n<text top=\"636\" left=\"401\" width=\"57\" height=\"18\" font=\"2\"> (pop) . </text>\n<text top=\"636\" left=\"459\" width=\"82\" height=\"18\" font=\"1\">depressão</text>\n<text top=\"636\" left=\"541\" width=\"5\" height=\"18\" font=\"2\"> </text>

A thick black curved arrow pointing downwards from the top right towards the bottom right.

"na milionésima parte de um grama\n (pop) , \n**micrograma**\n \nà volta da boca\n (pop) , \n**perioral**\n \nà volta da órbita\n (pop) , \n**periorbital**\n \nà volta dos vasos sanguíneos\n (pop) , \n**perivascular**\n \nabaixamento, abatimento, prostração\n (pop) , \n**depressão**\n \n**abcesso**\n , \n**nabcesso**, tumor\n (pop) \n**abcesso**, tumor\n (pop) , \n**abcesso**\n \n**abcesso**; acumulação de pus\n (pop) , \n**empiema**\n \n**abdómen**\n , \n**barreira**, ventre\n (pop) \n**abdominal**\n , \n**entral**\n (pop) \n**aberrante**\n , \n**anormal**\n (pop) \n**abertura**; orifício\n (pop) , \n**perfuração**\n \n**nablação**\n (pop) , \n**extracção**\n \n**nablação** dos órgãos sexuais, capaço, eviração, emasculação\n (pop) , \n**castração**\n \n**nabocamento**\n (pop) , \n**anastomose**\n \n**nabortamento**, desmancho\n (pop) , \n**aborto**\n \n**aborto**\n , \n**nabortamento**, desmancho\n (pop) \n\n\n**abrupto**\n , \n**repentino**, brusco\n (pop) \n**absorção**\n , \n**nabsorvimento**, absorvência\n (pop) \n**nabsorção** de água e de solutos por células vivas\n (pop) , \n**ressorção**\n \n**nabsorvimento**, absorvência\n (pop) , \n**absorção**\n \n**abstinência**\n , \n**njejum**\n (pop) \n**nabstracção**, suposição\n (pop) , \n**hipótese**\n \n**nacariase sarcóptica**, sarna\n (pop) , \n**escabiose**\n \n**acatisia**\n , \n**nincapacidade em permanecer sentado**\n (pop) \n**nacção** de certos corpos sobre outros\n (pop) , \n**catálise**\n \n**nacção** de enferrujar\n (pop) , \n**oxidação**\n \n**nacção** de inalar, extracção de líquidos ou gases\n (pop) , \n**aspiraçao**\n \n**nacção** de urinar, urinação\n (pop) , \n**micção**\n \n**naceleração** da frequência cardíaca\n (pop) , \n**taquicardia**\n \n**necessário**, anexo, indesejável\n (pop) , \n**secundário**\n \n**nacidade**, azedume\n (pop) , \n**acidez**\n \n**accidental**\n , \n**npor acaso, sem importância**\n (pop) \n**acidez**\n \n**nacidade**,



Tratamento dos dados

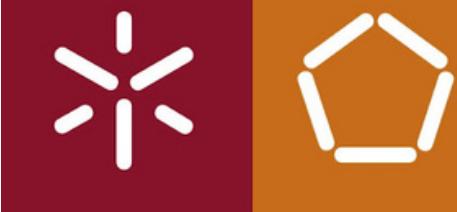
Glossário de Termos Médicos Técnicos e Populares.pdf

4. Limpeza de dados

- Remoção dos termos, parágrafos, vírgulas e excesso de espaços em branco

```
terms_re = r"<b>(.*)?</b>"
```

```
processed_content = re.sub(terms_re, "", content)
processed_content= re.sub(r"\n+", " ",
processed_content)
processed_content = re.sub(r"\s+,", "", 
processed_content)
processed_content = re.sub(r"\s{2,}", " ", 
processed_content)
```



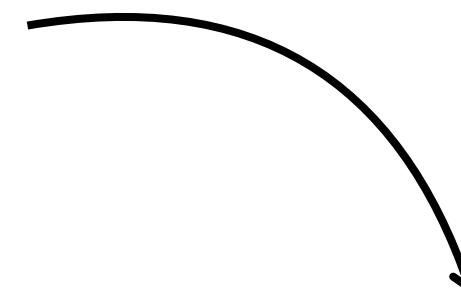
Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

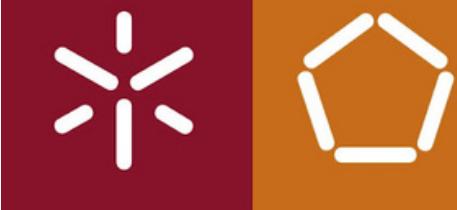
4. Limpeza de dados

- Remoção dos termos, parágrafos, vírgulas e excesso de espaços em branco

```
"\na milionésima parte de um grama\n (pop) , \nmicrograma\n \nà volta da boca\n (pop) , \nperioral\n \nà volta da órbita\n (pop) , \nperiorbital\n \nà volta dos vasos sanguíneos\n (pop) , \nperivascular\n \nabaixamento, abatimento, prostração\n (pop) , \ndepressão\n \nabcesso\n , \nnabcesso, tumor\n (pop) \nnabcesso, tumor\n (pop) , \nabcesso\n \nnabcesso; acumulação de pus\n (pop) , \nempíema\n \nabdómen\n , \nbarriga, ventre\n (pop) \nabdominal\n , \nventral\n (pop) \naberrante\n , \nanormal\n (pop) \nabertura; orifício\n (pop) , \nperfuração\n \nnablação\n (pop) , \nextracção\n \nnablação dos órgãos sexuais, capaço, eviração, emasculação\n (pop) , \ncastração\n \nabocamento\n (pop) , \nanastomose\n \nnabortamento, desmancho\n (pop) , \naborto\n \naberto\n , \nnabortamento, desmancho\n (pop) \nabrupto\n , \nrepentino, brusco\n (pop) \nabsorção\n , \nnabsorvimento, absorvência\n (pop) \nnabsorção de água e de solutos por células vivas\n (pop) , \nressorção\n \nnabsorvimento, absorvência\n (pop) , \nabsorção\n \nabstinência\n , \njejum\n (pop) \nnabstracção, suposição\n (pop) , \nhipótese\n \nnacriase sarcóptica, sarna\n (pop) , \nescabiose\n \nacatisia\n , \nincapacidade em permanecer sentado\n (pop) \nnacção de certos corpos sobre outros\n (pop) , \ncatálise\n \nnacção de enferrujar\n (pop) , \noxidação\n \nnacção de inalar, extracção de líquidos ou gases\n (pop) , \naspiração\n \nnacção de urinar, urinação\n (pop) , \nmicção\n \nnaceleração da frequência cardíaca\n (pop) , \ntaquicardia\n \nnecessório, anexo, indesejável\n (pop) , \nsecundário\n \nnacidade, azedume\n (pop) , \nacidez\n \naccidental\n , \npor acaso, sem importância\n (pop) \nacidez\n , \nnacidade,
```



1 " a milionésima parte de um grama (pop) à volta da boca (pop) à volta dos vasos sanguíneos (pop) abaixamento, abatimento, prostração (pop) abcesso, tumor (pop) abcesso, tumor (pop) abcesso; acumulação de pus (pop) barriga, ventre (pop) ventral (pop) anormal (pop) abertura; orifício (pop) ablação (pop) ablação dos órgãos sexuais, capaço, eviração, emasculação (pop) abocamento (pop) abortamento, desmancho (pop) abortamento, desmancho (pop) repentina, brusco (pop) absorvimento, absorvência (pop) absorção de água e de solutos por células vivas (pop) absorvimento, absorvência (pop) jejum (pop) abstracção, suposição (pop) acariase sarcóptica, sarna (pop) incapacidade em permanecer sentado (pop) acção de certos corpos sobre outros (pop) acção de enferrujar (pop) acção de inalar, extracção de líquidos ou gases (pop) acção de urinar, urinação (pop) aceleração da frequência cardíaca (pop) acessório, anexo, indesejável (pop) acidez, azedume (pop) por acaso, sem importância (pop) acidez, azedume (pop) ácido animado (pop) alteração do equilíbrio ácido básico do sangue e líquidos teciduais (pop) ausência de movimento, acinese (pop) espinha (pop) adaptação (pop) acomodação (pop) acompanhante (pop) acostumado (pop) acresentar um solvente a um medicamento em pó; tipo de regeneração de uma superfície lesionada; retorno à forma líquida do soro ou plasma sanguíneo previamente dessecado (pop) cor azulada das extremidades (mãos-pés) (pop) hormônio adreno-corticotrófico, corticotrofina (pop) activação de uma droga por outra (pop) activação de uma droga por outra (pop) activado por adrenalina (pop) estimular ou acelerar, reactivar (pop) actividade (pop) actividade (pop) actividade excessiva da tiróide (pop) actividade insuficiente da tiróide (pop) reactivo, eficaz (pop) activo com vários grupos de germes (pop) acto de relembrar os antecedentes do doente (pop) factor II da coagulação sanguínea; trombogénio (pop) mistura de glicose e frutose (pop) agudez(a) (pop) acumulamento (pop) acumulação de ureia no sangue; urinemia (pop) acumulação excessiva de glóbulos gordos nos tecidos (pop) acumulamento (pop) acumulativo (pop) acomodação (pop) adaptação (pop) caroço inflamado, inflamação das glândulas (pop) tumor benigno (pop) apropriado, próprio (pop) adesão, união viciosa das superfícies orgânicas (pop) adesão, união viciosa das superfícies orgânicas (pop) complementar (pop) produto



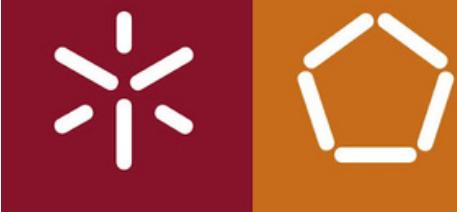
Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

4. Marcações

```
processed_content = re.sub(r"\(pop\)", "@@", processed_content)
```

" a milionésima parte de um grama @@ à volta da boca @@ à volta da órbita @@ à volta dos vasos sanguíneos @@ abaixamento, abatimento, prostração @@ abcesso, tumor @@ abcesso, tumor @@ abcesso; acumulação de pus @@ barriga, ventre @@ ventral @@ anormal @@ abertura; orifício @@ ablcação @@ ablcação dos órgãos sexuais, capação, eviração, emasculação @@ abocamento @@ abortamento, desmancho @@ abortamento, desmancho @@ repentina, brusco @@ absorvimento, absorvência @@ absorção de água e de solutos por células vivas @@ absorvimento, absorvência @@ jejum @@ abstracção, suposição @@ acariase sarcóptica, sarna @@ incapacidade em permanecer sentado @@ acção de certos corpos sobre outros @@ acção de enferrujar @@ acção de inalar, extracção de líquidos ou gases @@ acção de urinar, urinação @@ aceleração da frequência cardíaca @@ acessório, anexo, indesejável @@ acidade, azedume @@ por acaso, sem importânci @@ acidade, azedume @@ ácido animado @@ alteração do equilíbrio ácido básico do sangue e líquidos teciduais @@ auséncia de movimento, acinese @@ espinha @@ adaptação @@ acomodação @@ acompanhante @@ acostumado @@ acrescentar um solvente a um medicamento em pó; tipo de regeneração de uma superfície lesionada; retorno à forma líquida do soro ou plasma sanguíneo previamente dessecado @@ cor azulada das extremidades (mãos-pés) @@ hormônio adreno-corticotrófico, corticotrofina @@ activação de uma droga por outra @@ activação de uma droga por outra @@ activado por adrenalina @@ estimular ou acelerar, reactivar @@ actividade @@ actividade @@ actividade excessiva da tiróide @@ actividade insuficiente da tiróide @@ reactivo, eficaz @@ activo com vários grupos de germes @@ acto de relembrar os antecedentes do doente @@ factor II da coagulação sanguínea; trombogénio @@ mistura de glicose e frutose @@ agudez(a) @@ acumulamento @@ acumulação de ureia no sangue; urinemia @@ acumulação excessiva de glóbulos gordos nos tecidos @@ acumulamento @@ acumulativo @@ acomodação @@ adaptação @@ caroço inflamado, inflamação das glândulas @@ tumor benigno @@ apropriado, próprio @@ adesão, união viciosa das superfícies orgânicas @@ adesão, união viciosa das superfícies orgânicas @@ complementar @@ produto acrescentado @@ substância que contribui para a auxiliar @@ administração de medicamentos como sedativos antes de uma anestesia geral @@ administração de um antigénio para induzir uma resposta imunitária @@ administração de um medicamento ou emprego de uma medida física @@ administraçãogota-a-gota (solução oftálmica) @@ administrado por via não oral @@ jovem @@ activado por adrenalina @@ substância que inibe a resposta à adrenalina @@ penetração pela pele ou pelas mucosas, consumo @@ que aperta os tecidos, secura da pele @@ adulto @@ adulto @@ aeragem, arejamento, aeração @@ que vive no ar @@ engolir ar @@ solução de um produto destinado à inalação @@ perda da capacidade de comunicar @@ afecção do fígado @@ afinação, ajustamento @@ parentesco, analogia, semelhança @@ afundamento ou queda (ex.: útero, recto, cordão umbilical) @@ falta ou diminuição da secreção do leite @@ sem gamoglobinas @@ substância activa @@ agente intermediário, substância química que transmite algo @@ agente que destrói ou dissolve a mucina @@ agente que dissolve a parte córnea da pele @@ agente que promove a secreção urinária do ácido úrico @@ agente que reduz a tensão @@ desassossego, inquietação, excitação @@ músculo que participa com outro num movimento; medicamento que estimula células de maneira natural @@ inexistência de glóbulos no sangue, doença de Pfeiffer @@ agravamento, pioramento, piora @@ agravamento, pioramento, piora @@ agregado, associação @@ agregado, associação @@



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

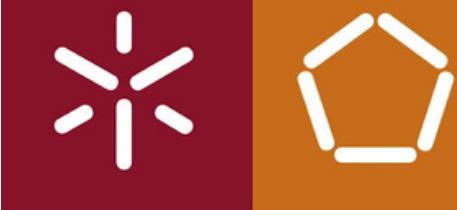
4. Extração da informação relevante

```
terms = re.findall(terms_re, content)
terms = [termo.strip() for termo in terms]
```

```
design = re.findall(r"(?:@(?:\s|,|X)|^)(.*?)(?:@)", processed_content)
design = [desc.strip() for desc in design]
```

```
glossario_dict = dict(zip(terms, design))
```

```
for termo, descricao in zip(terms, design):
    if termo in glossario_dict:
        if isinstance(glossario_dict[termo], list):
            if descricao not in glossario_dict[termo]:
                glossario_dict[termo].append(descricao)
        else:
            if descricao != glossario_dict[termo]:
                glossario_dict[termo] = [glossario_dict[termo], descricao]
    else:
        glossario_dict[termo] = descricao
```



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

5. Intervenções manuais

- Junção de termos separados

- pré-medicação

```
331 <text top="350" left="777" width="31" height="18" font="1"><b>pré-</b></text>
332 <text top="369" left="128" width="85" height="18" font="1"><b>medicação</b></text>
```

- infeção cruzada

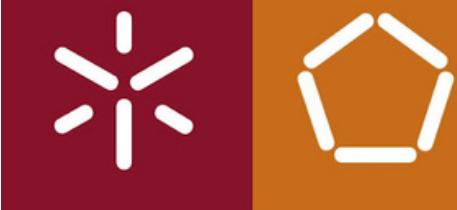
```
2773 <text top="873" left="741" width="67" height="18" font="1"><b>infecção</b></text>
2774 <text top="892" left="128" width="62" height="18" font="1"><b>cruzada</b></text>
```

- efeito colateral

```
4324 <text top="882" left="764" width="44" height="18" font="1"><b>efeito</b></text>
4325 <text top="901" left="128" width="68" height="18" font="1"><b>colateral</b></text>
```

- tremor intencional

```
14419 <text top="530" left="755" width="52" height="18" font="1"><b>tremor</b></text>
14420 <text top="550" left="128" width="87" height="18" font="1"><b>intencional</b></text>
```



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

5. Intervenções manuais

- Correção de erros ortográficos
 - protrombina

```
<text top="674" left="128" width="346" height="18" font="5"><i>actor II da coagulação sanguínea; trombogénio</i></text>
<text top="795" left="239" width="351" height="18" font="5"><i>factor II da coagulação sanguínea; trombogénio</i></text>
```

- pré-medicação

```
<text top="350" left="128" width="594" height="18" font="5"><i>administração de medicamentos como sedativos antes de uma anestesia gera </i></text>
<text top="552" left="260" width="546" height="18" font="5"><i>administração de medicamentos como sedativos antes de uma anestesia</i></text>
<text top="571" left="128" width="37" height="18" font="5"><i>geral</i></text>
```

- Remoção de símbolos resultantes da utilização de aspas
 - dooping

```
<text top="998" left="188" width="70" height="18" font="5"><i>&#34;dooping&#34;</i></text>
```



Tratamento dos dados

Glossário de Termos Médicos Técnicos e Populares.pdf

RESULTADO